Biostatistics (2003), **4**, 2, *pp.* 265–278
Printed in Great Britain

Insights on bias and information in group-level studies

LIANNE SHEPPARD
Departments of Biostatistics and Environmental Health, Box 357232, University of Washington, Seattle, WA 98195-7232, USA
sheppard@u.washington.edu

SUMMARY

Ecological and aggregate data studies are examples of group-level studies. Even though the link between the predictors and outcomes is not preserved in these studies, inference about individual-level exposure effects is often a goal. The disconnection between the level of inference and the level of analysis expands the array of potential biases that can invalidate the inference from group-level studies. While several sources of bias, specifically due to measurement error and confounding, may be more complex in group-level studies, two sources of bias, cross-level and model specification bias, are a direct consequence of the disconnection. With the goal of aligning inference from individual versus group-level studies, I discuss the interplay between exposure and study design. I specify the additional assumptions necessary for valid inference, specifically that the between- and within-group exposure effects are equal. Then cross-level inference is possible. However, all the information in the group-level analysis comes from between-group comparisons. Models where the group-level analysis provides even a small percentage of information about the within-group exposure effect are most susceptible to model specification bias. Model specification bias can be even more serious when the group-level model isn't derived from an individual-level model.

Keywords: Aggregate data studies; Cross-level inference; Ecological studies; Relative risk.

1. INTRODUCTION

Ecological studies are frequently maligned but often used in epidemiology because of their relatively low cost and the availability of data (Morgenstern, 1982; Greenland, 1992). For example, in cancer epidemiology, ecological studies of diet and cancer relate routinely collected food consumption data to cancer incidence rates across geographic areas. The low cost and ability of these studies to capture dietary contrasts that are more extreme than those present within areas make them appealing (Prentice and Sheppard, 1990; Prentice *et al.*, 1988). However, their use is controversial (e.g. Willett and Stampfer (1990)). In contrast, ecological studies are common in air pollution epidemiology. The most common type of study, the time series study, is an ecological study. Time series studies relate daily population health outcomes to daily air pollution exposure measurements (e.g. Sheppard *et al.* (1999)). Since they are based on routinely collected data, they are inexpensive to conduct. Unlike other ecological studies, the air pollution time series studies are not viewed as inherently flawed. They form an important part of the evidence upon which air quality standards are based (e.g. EPA (1996)).

Since groups are the unit of analysis, ecological studies are group-level studies. There is no linkage in the data between individual exposures and outcomes in these studies. Inference from ecological studies

can be either at the level of the analysis, i.e. ecological inference, or at the individual level, called biologic inference (Morgenstern, 1998). Often biologic inference is the goal of an ecological study even though the data are at the group level. This is a form of cross-level inference. The disconnection between the level of the analysis and the level of inference means additional sources of bias can enter beyond those ordinarily of concern with any observational study.

This paper will focus on biologic inference, specifically with respect to the exposure effect estimate(s), from group-level studies. Because of the potential biases, many scientists have counseled against the use of ecologic studies for inference (e.g. Piantadosi *et al.* (1988)). However, given the compelling need to make the best possible use of available data to benefit public health, there will continue to be interest in inferring individual-level risks from group-level data. Practical limitations in exposure assessment in individual-level studies also direct more attention to group-level studies. I suggest that a constructive response to the natural desire to use group-level data for biologic inference in public health is to develop methods that will draw valid inference while making the best use of available data. This, combined with active recognition of the strengths and weaknesses in a specific application, will facilitate progress in elucidating the health effects of environmental exposures.

Because of their history and data constraints, ecological models are often mis-specified. In many settings model mis-specification alone is sufficient to ensure ecological study parameters cannot be used for biologic inference. The first step in any group-level analysis is to align model specification with individual-level studies. Beyond this, group-level studies require additional assumptions. Specifically, because biologic inference is cross-level, one must consider the potential for cross-level bias.

The study of dietary fat and breast cancer is a good motiviating example. Early evidence of the association between dietary fat and cancer came from international ecological studies (Armstrong and Doll, 1975; Doll and Peto, 1981). Within-country prospective cohort studies are considered the most inferentially sound of all observational studes (e.g. Willett (2001), Zock (2001)), but many have failed to replicate this association for breast cancer (e.g. Homles *et al.* (1999)). However, measurement error bias is a major challenge in all observational studies of diet. In fact, ecological studies showing an association between fat intake and breast cancer incidence can be shown to be broadly consistent with many analytic epidemiology studies when measurement error bias is incorporated (Prentice and Sheppard, 1990). Furthermore, because population rather than individual estimates of dietary intake are used, international comparisons are not subject to the same exposure measurement error bias that plagues traditional case-control and cohort studies within populations (Prentice and Sheppard, 1995; Sheppard and Prentice, 1995). It is likely that the prevailing preference of analytic epidemiological studies over ecological studies for the study of diet and cancer heavily discounts the importance of measurement error as a dominant source of bias while simultaneously emphasizing other biases that occur in group-level studies. Regardless, considerable skepticism remains regarding the interpretability of ecological studies of fat and cancer, particularly in view of their reliance on readily available food consumption data and their limited ability to control for confounding. Even less is understood about the role of types of effects, proper model specification, and the exposure information available in the analysis. These are the main topics of this paper. I begin by discussing a framework for classifying designs based on grouping in the exposures and/or outcomes. This conceptually links individual and group-level designs.

2. GROUPING IN EXPOSURE VERSUS ANALYSIS

In multi-popoulation ecologic studies, events are tallied over a fixed case accession period. In the cancer incidence application it is appropriate to assume these are rare events. This assumption also is appropriate for many other epidemiologic investigations of health effects of environmental exposures. Outcomes may range from total mortality, incidence of cancer or a chronic disease, to medical service use such as emergency department visits or aid car responses for cardiovascualar emergencies.

For the model at the individual level, assume $Y_{ki} = 1$ indicates a disease event of interest (0 otherwise) for individual $i = 1, \ldots, n_k$ within area A_k, $k = 1, \ldots, K$ during a fixed time period. Under a rare disease assumption, a plausible relative risk model is

$$\mu_{ki}(I) = \mathrm{E}(\mathrm{E}(Y_{ki}|v_k, \mathbf{x}_{ki})) = \mathrm{E}(\exp(\alpha + \mathbf{x}_{ki}^T\boldsymbol{\beta} + v_k)) \tag{2.1}$$

where the outer expectation is with respect to v_k, an area-specific random variable with $\mathrm{E}(e^{v_k}) = 1$ and $\mathrm{var}(e^{v_k}) = \sigma_v^2$, and $\mathbf{x}_{ki}$ is a P-vector of exposures and confounders. The regression parameters $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta})$ are to be estimated where, specifically, components of $\boldsymbol{\beta}$ are exposures of scientific interest. I assume this individual model is valid and investigate the implications of modifications to it due to study design and data availability.

A useful conceptualization of grouping in epidemiology is to distinguish between grouping in exposures (or more generally any covariates) and study design and/or analysis. Table 1 depicts the two aspects of grouping. Each cell gives the corresponding type of study, the mean as defined in this section, and the equation number where the mean is defined. Exposures are grouped whenever a group mean exposure is substituted for an individual exposure. Departing from (2.1), the analysis would rely on substituting the group mean, $\overline{x}_{kp} = \sum_{i=1}^{n_k} x_{kip}/n_k$, for one or more of the $p = 1, \ldots, P$ covariates. For all covariates grouped, the model is

$$\mu_{ki}(S) = \mathrm{E}(\exp(\alpha_s + \overline{\mathbf{x}}_k^T\boldsymbol{\beta}_s + v_k)). \tag{2.2}$$

The subscript s indicates the parameters $\boldsymbol{\theta}_s$ may no longer equal $\boldsymbol{\theta}$ since the predictors have changed. This approach is often taken in multi-city air pollution epidemiology cohort studies. In such cases the air pollution exposure is a community-level average over time from one or more monitors while other exposure and confounding variables are available for each individual (Dockery *et al.*, 1993; Pope *et al.*, 1996). Kunzli and Tager (1997) call this the semi-individual design.

The alternate approach to grouping is to group the analysis. Here the linkage between covariates and outcomes is severed at the individual level and thus the analysis is conducted at the level of the group. The most direct transition from an individual to grouped data study is the aggregate data study (Prentice and Sheppard, 1995; Sheppard and Prentice, 1995). Building from (2.1), the aggregate model is

$$\mu_k(A) = \mathrm{E}(\mathrm{E}(\bar{Y}_k|\mathbf{x}_{ki}, v_k)) = \mathrm{E}(n_k^{-1}\sum_{i=1}^{n_k}\exp(\alpha_a + \mathbf{x}_{ki}^T\boldsymbol{\beta}_a + v_k)) \tag{2.3}$$

where $\bar{Y}_k = \sum_{i=1}^{n_k} Y_{ki}/n_k$. This is an example of a 'complete data' aggregate model where individual covariates are assumed to be available for all members in each group. In order to be realistic in application, an aggregate data study must rely on a covariate subsample from the population (Prentice and Sheppard, 1995). However, in order to address conceptual differences between ecological and aggregate studies, I will restrict attention in this paper to the complete data aggregate model. The subscript a on the aggregate model parameters allows for the possibility that they differ from the individual-level model parameters. A goal of this paper is to clarify when $\boldsymbol{\beta} = \boldsymbol{\beta}_a$.

The ecological model follows directly from the aggregate model by subsitituting group mean exposures for individual exposures. Here the ecological model retains the same log–linear form specified at the individual level:

$$\mu_k(E) = \exp(\alpha_e + \bar{\mathbf{x}}_k^T\boldsymbol{\beta}_e). \tag{2.4}$$

The 'e' subscript emphasizes the potential change in parameter interpretation for $\boldsymbol{\theta}_e$. Consistent with the majority of the ecological study literature, there is no random effect (v_k) in this model. Linkage of this

Table 1. *Interplay between exposure and design/analysis*

	Design/Analysis Level	
Exposure	Individual	Group
Individual	Individual $\mu_{ki}(I)$, (2.1)	Aggregate Data $\mu_k(A)$, (2.3)
Group	Semi-Individual $\mu_{ki}(S)$, (2.2)	Ecologic $\mu_k(E)$, (2.4)

model with the disease mapping literature would include both v_k as well as the spatially dependent random effect u_k (Besag *et al.*, 1991). This particular ecological model is mis-specified because it substitutes the average covariate values into the relative risk function. (Note that there is no mis-specification for the special case of a linear relative risk model with a single covariate.)

Often ecological models are specified without direct alignment with the presumed individual-level relationship (e.g. Greenland and Robins (1994), Greenland (2001)). In an ecological study the norm is to begin by specifying a group-level model with little or no regard for the individual relationships. Published studies commonly specify linear models, regardless of the hypothesized effect of exposure on the disease outcome at the individual level. One important reason why ecological studies often do not estimate the parameters of interest stems from this fundamental discrepancy between their set-up and intended interpretation. So even though it is possible to specify ecological models with the same functional form as an individual-level study model, ecological studies in the published literature often suffer from both types of model mis-specification: model form and grouped covariates. Guthrie and Sheppard (2001) show through direct simulation of the examples given in Greenland and Robins (1994) that the aggregate data study is free from many of the biases due to model mis-specification in ecological studies. In Section 3.5 I give further detail on model specification bias in ecological versus aggregate studies.

In this paper I am concerned with deeper understanding of inference in ecological studies relative to that possible in individual-level studies. I am interested in biologic inference of exposure effect parameters, i.e. those components of $\boldsymbol{\beta}$ associated with environmental exposures. I assume the exposures to individuals will vary over individuals, both within and between groups. First I address a key assumption that is necessary in order to make valid biologic inference from group-level studies: the assumption of no contextual effects (Section 3.2). Then I address the information available to an individual-level versus the equivalent group-level analysis (Section 3.3). This is most readily done by comparing individual to aggregate studies. Not only does this clarify the sources of information for a group-level study, but it also clarifies situations when there is greater potential for specification bias. This leads into a discussion of specification bias due to substituting group mean exposures for the distribution of within-group exposure as is done when one moves from aggregate to ecological studies (Section 3.5).

3. LINKING INDIVIDUAL-LEVEL WITH GROUP-LEVEL STUDIES

3.1 *Preliminaries*

The individual (2.1), aggregate (2.3), and ecological (2.4) unconditional means and variances are, respectively,

$$\begin{aligned}
\mu_{ki}(I) &= \mathrm{E}(Y_{ki}) = \exp(\alpha + \mathbf{x}_{ki}\boldsymbol{\beta}) & \mathbf{V}_k(I) &= \Delta_k + \sigma^2 \boldsymbol{\mu}_k(I)\boldsymbol{\mu}_k^T(I) \\
\mu_k(A) &= \mathrm{E}(\bar{Y}_k) = n_k^{-1}\textstyle\sum_{i=1}^{n_k} \exp(\alpha + \mathbf{x}_{ki}\boldsymbol{\beta}) & V_k(A) &= \sigma^2(\mu_k(A)^2 - \phi_k/n_k) + (\mu_k(A) - \phi_k)/n_k \\
\mu_k(E) &= \exp(\alpha_e + \bar{\mathbf{x}}_k\boldsymbol{\beta}_e) & V_k(E) &= \mu_k(E)(1 - \mu_k(E))/n_k
\end{aligned}$$

where $\boldsymbol{\mu}_k(I) = (\mu_{k1}(I), \ldots, \mu_{kn_k}(I))$, $\Delta_k = \text{diag}(\mu_{ki}(I)[1-(1+\sigma^2)]\mu_{ki}(I))$, and $\phi_k = \sum_{i=1}^{n_k} \mu_{ki}^2(I)$. Consistent with the level of analysis, the aggregate and ecological models yield a single mean for each group while the individual model produces a vector of means in each group. Likewise the variances for the ecological and aggregate models are scalar while the individual model variance is a matrix.

3.2 *Cross-level inference*

There are many potential sources of bias in ecological studies. These are variously described in the literature by terms such as ecological fallacy, aggregation bias, specification bias, and cross-level bias (e.g. Morgenstern and Thomas (1993), Morgenstern (1982), Greenland and Morgenstern (1989), Langbein and Lichtman (1978)). Inconsistent and incomplete definitions of these terms have contributed to the difficulty of comprehending the pitfalls and benefits of group-level studies.

The original definition of cross-level bias appears to be due to Firebaugh (1978, p. 560) where he states that 'Cross-level bias is absent when, and only when, $\beta_2 = 0$ in the structural equation $Y = a + \beta_1 X_1 + \beta_2 \bar{X}_1 + e$.' I adopt a modified version of this definition below. Morgenstern (1982) introduced Firebaugh's definition into the public health literature. He went on to define cross-level bias as the sum of aggregation and specification biases, terms that weren't also written mathematically. In later contributions, Greenland and Morgenstern (1989) used cross-level bias as a synonym for ecological bias while Richardson *et al.* (1987) used the term to describe uncontrolled between-group confounding.

The concept of cross-level bias is closely related to the distinction between between- and within-group covariate effects as discussed in the longitudinal data literature (e.g. Neuhaus and Kalbfleisch (1998)). Diggle et al. (1994, p.23ff) use slightly different terminology but address the same issue when they discuss the difference in interpretation between cross-sectional (between-cluster) and longitudinal (within-cluster) effects. Much earlier, Scott and Holt (1982) noted the importance of assuring that the two slopes β_B and β_W are equal in the model

$$Y_{ki} = \alpha + \bar{x}_k \beta_B + (x_{ki} - \bar{x}_k)\beta_W + \epsilon_{ij}.$$

This is a prerequisite to their work addressing the efficiency of least-squares estimates in clustered data settings. With the goal of removing uncontrolled confounding in longitudinal studies, Palta *et al.* (1994) stressed the importance of controlling for group effects, period effects, and other potential confounders that vary differently between and within individuals in longitudinal studies. They suggest checking for discrepancies and testing the coefficients of $\bar{\mathbf{x}}_k$ in a regression model as a way of identifying possible omitted variables. Likewise in the psychology literature, Schwartz and Stone (1998) advocate separating pure within-cluster from between-cluster effects in the analysis of 'ecological momentary assessment' data, the term they use for repeated observational data.

With regard to the group-level analysis setting, suppose a general individual-level model is defined as

$$g(\text{E}(Y_{ki}|\bar{\mathbf{x}}_k, \mathbf{x}_{ki})) = \alpha + \bar{\mathbf{x}}_k \boldsymbol{\beta}_B + (\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)\boldsymbol{\beta}_W,$$

where $\mathbf{x}$ is a vector of exposure and confounder variables. I have explicitly separated the predictors into components that vary within versus between groups. As in the longitudinal data setting, this notation recognizes the potential difference in many studies of between-group and within-group covariate effects. In epidemiology typically we are interested in the effect of an exposure on an individual's risk of a disease outcome. The parameter of interest for inference on individuals (biologic inference) is β_W. As I demonstrate in the next section, most of the information provided in a group-level analysis is for β_B. We incur cross-level bias in group-level studies whenever we assume $\beta_W = \beta_B$ when in fact $\beta_W \neq \beta_B$. I suggest that when the goal is biologic inference, a group-level analysis is appropriate *only* in applications where one is comfortable hypothesizing that $\beta_W = \beta_B$.

Writing Firebaugh's model using the above notation gives

$$g(\mathrm{E}(Y_{ki}|x_{ki}, \bar{x}_k)) = \alpha + x_{ki}\beta_1 + \bar{x}_k\beta_2.$$

Since $\beta_W = \beta_1$ and $\beta_B = \beta_1 + \beta_2$, Firebaugh's assumption that $\beta_2 = 0$ is equivalent to assuming $\beta_B = \beta_W$. The key criterion necessary to ensure no cross-level bias is that there is no additional effect due to the overall exposure level on the group above its effect on an individual. Firebaugh called these contextual effects (Firebaugh, 1978; Hammond, 1973; Greenland, 2001, 2002). The most natural examples of contextual variables in epidemiology are socially defined exposures. For instance, it is plausible to assume that the overall deprivation in an area would have an additional contextual effect on health beyond the effect of an individual's degree of deprivation. As another example, the media can have an effect both on individuals and on the group by influencing the social norms in an area.

Note the distinction between parameters and estimates here. In many applications it may be reasonable scientifically to assume that the underlying parameters are equal. Regardless, their estimates may differ. The discrepancies could be caused by uncontrolled confounding or exposure measurement error. These biases will operate differently on β_W and β_B and can be investigated directly in a multi-population cohort study because it can estimate both parameters. For instance, it is well known that dietary data have a significant amount of measurement error (Prentice, 1996) so that estimates of β_W are typically attenuated. However, the error distribution may be reasonably assumed to be constant across cohorts. Then the dietary measurement error wouldn't bias an estimate of β_B. Thus, assuming $\beta_B = \beta_W$, measurement error bias will still result in a difference between $\hat{\beta_B}$ and $\hat{\beta_W}$.

Biases can dominate one level over another due to differences in exposure, confounder, and measurement error distributions across levels. In all studies, judgment about the equivalence of the underlying within- and between-group parameters must be made on a scientific basis. Individual-level studies offer the opportunity to test directly for differences in them, while group-level studies do not. However, even for individual-level studies, in observational settings one cannot rule out the possibility that observed differences in $\hat{\beta_B}$ and $\hat{\beta_W}$ are due to bias in the estimates as opposed to underlying differences in the parameters.

3.3 *Information sources for health effect parameter estimation*

This section shows how much information is available in individual-level and aggregate data studies to estimate β_W and β_B. Note that the ecological model, e.g. $\mu_k(E) = \exp(\alpha + \bar{\mathbf{x}}_k\beta)$, does not use any within-group information from the covariate distribution. This contrasts with the aggregate study that uses some within-group information by averaging over functions of all covariates. The important distinction between ecological and aggregate studies is specification bias (see Section 3.5).

Sheppard and Prentice (1995) give the information for the parameter $\boldsymbol{\beta}$ for individual-level and aggregate models. It is calculated using the general formula for the information of $\boldsymbol{\beta}$ alone, $I(\boldsymbol{\beta}) = I_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\theta}) - I_{\boldsymbol{\beta}\boldsymbol{\alpha}}(\boldsymbol{\theta})I^{-1}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}(\boldsymbol{\theta})I_{\boldsymbol{\beta}}(\boldsymbol{\theta})$ for

$$I(\boldsymbol{\theta}) = \begin{bmatrix} I_{\boldsymbol{\alpha}\boldsymbol{\alpha}}(\boldsymbol{\theta}) & I_{\boldsymbol{\alpha}\boldsymbol{\beta}}(\boldsymbol{\theta}) \\ I_{\boldsymbol{\beta}\boldsymbol{\alpha}}(\boldsymbol{\theta}) & I_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\theta}) \end{bmatrix}. \tag{3.1}$$

The resulting information is written quite generally in terms of sums of squared differences in exposures and weighted average exposures. I extend this to understand sources of information by explicitly separating the within- and between-group covariates and effects. The individual (2.1) and aggregate (2.3) models become

$$\mu_{ki}(I) = \exp(\alpha + \bar{\mathbf{x}}_k\boldsymbol{\beta}_B + (\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)\boldsymbol{\beta}_W) \tag{3.2}$$

and

$$\mu_k(A) = n_k^{-1} \sum_{i=1}^{n_k} \exp(\alpha + \bar{\mathbf{x}}_k \boldsymbol{\beta}_B + (\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)\boldsymbol{\beta}_W). \tag{3.3}$$

This allows the two vectors of parameters, $\boldsymbol{\beta}_W$ and $\boldsymbol{\beta}_B$, to be distinct. Suppose there is only a single exposure so $\mathbf{x} = (x_{ki} - \bar{x}_k, \bar{x}_k)$ and $\boldsymbol{\beta} = (\beta_W, \beta_B)$. Then the information equations for the individual and aggregate models are, respectively,

$$\begin{aligned} I_I(\boldsymbol{\beta}) = &\sum_{k=1}^{K}\sum_{i=1}^{n_k} d_{ki} \begin{bmatrix} (x_{ki} - \bar{x}_k(d))^2 & 0 \\ 0 & 0 \end{bmatrix} + \sum_{k=1}^{K} c_{k.} \begin{bmatrix} g_k(c)^2 & (\bar{x}_k - \overline{\bar{x}(c)})g_k(c) \\ g_k(c)(\bar{x}_k - \overline{\bar{x}(c)}) & (\bar{x}_k - \overline{\bar{x}(c)})^2 \end{bmatrix}, \\ I_A(\boldsymbol{\beta}) = &\sum_{k=1}^{K} e_{k.} \begin{bmatrix} g_k(e)^2 & (\bar{x}_k - \overline{\bar{x}(e)})g_k(e) \\ g_k(e)(\bar{x}_k - \overline{\bar{x}(e)}) & (\bar{x}_k - \overline{\bar{x}(e)})^2 \end{bmatrix}, \end{aligned} \tag{3.4}$$

where $g_k(c) = (\bar{x}_k(c) - \overline{x(c)}) - (\bar{x}_k - \overline{\bar{x}(c)})$, $\overline{\bar{x}(c)} = \sum_k c_{k.}\bar{x}_k / \sum_k c_{k.}$, and the notation is identical for $g_k(e)$. Furthermore, for the individual model $\bar{\mathbf{x}}_k(d) = \sum_i d_{ki}\mathbf{x}_{ki} / \sum_i d_{ki} = \sum_i c_{ki}\mathbf{x}_{ki} / \sum_i c_{ki} = \bar{\mathbf{x}}_k(c)$, and $\overline{\mathbf{x}(c)} = \sum_k c_{k.}\bar{\mathbf{x}}_k(c) / \sum_k c_{k.}$ where $d_{ki} = \mu_{ki}(I)[1 - (1+\sigma^2)\mu_{ki}(I)]^{-1}$, $\cdot$ denotes summation, and $c_{ki} = d_{ki}(1+\sigma^2 d_{k.})^{-1}$. At the aggregate level, $e_{ki} = \mu_{ki}(I)\mu_k(A)/V_k(A)n_k$, $\bar{\mathbf{x}}_k(e) = \sum_i e_{ki}\mathbf{x}_{ki} / \sum_i e_{ki}$, and $\overline{\mathbf{x}(e)} = \sum_k e_{k.}\bar{\mathbf{x}}_k(e) / \sum_k e_{k.}$.

Note that the individual-level model information, $I_I(\boldsymbol{\beta})$, has two components, one that sums over individuals and groups (the 'within groups component') and one that sums only over groups (the 'between groups component'). In contrast, the aggregate model information, $I_A(\boldsymbol{\beta})$, has only the between groups component. Within each component the information is given for (β_W, β_B). For the individual-level model β_W uses information from both components, while β_B uses information only from the between-group component. All information about β_B comes from summing over groups. Likewise, even though the aggregate model lacks the within-group component, it still contributes some information about β_W from the between-group component. The information about β_W from combining groups is obtained from $g_k(c)^2$ or $g_k(e)^2$, squared differences in weighted mean differences from weighted grand means. One can expect these differences will typically be small, particularly for covariates with symmetric distributions within groups. These terms will be exactly 0 whenever $\boldsymbol{\beta} = 0$ (since $g_k(\cdot) = 0$). Finally, it can be shown that the between-group components are equal in both models when $\mu_{ki}(I)$ is very small.

This exercise gives the information in an aggregate versus individual model when the parameters β_W and β_B remain distinct, regardless of whether or not they are equal. It points up two different issues regarding the goal of biologic inference. First, the information in the data about the exposure effect parameters β_W and β_B can be partitioned into between-group and within-group components. The within-group component does not provide any information for estimation of the between-group exposure effect parameter, β_B. In contrast, the between-group component provides information about both β_B and β_W. However, we expect the information about β_W in the between-group component will typically be small. Second, the parameter of interest for biologic inference is always β_W. As we expect that there is very little information about this parameter in an aggregate study, biologic inference from an aggregate (or more generally any group-level) study will be valid only when $\beta_W = \beta_B$. Biologic inference from an aggregate study is cross-level because even when we parametrize the model so $\beta = \beta_W = \beta_B$, the information in the data allow us predominantly to estimate β_B while the scientifically justified parameter for biologic inference is β_W.

Table 2. *Sources of information in individual-level and aggregate models with between and within-group exposure effect parameters*

$\frac{\sigma_B^2}{\sigma_T^2}$	Fraction of Total Individual Information							
	Within Area Component				Between Area Component			
	$I_I(\beta_W)$	$I_A(\beta_W)$	$I_I(\beta_B)$	$I_A(\beta_B)$	$I_I(\beta_W)$	$I_A(\beta_W)$	$I_I(\beta_B)$	$I_A(\beta_B)$
	$X \sim lognormal$							
0.22	0.94	–	0	–	0.03	0.03	0.03	0.03
0.35	0.87	–	0	–	0.02	0.02	0.12	0.12
0.55	0.69	–	0	–	0.006	0.007	0.31	0.32
0.73	0.47	–	0	–	0.002	0.002	0.52	0.53
0.92	0.16	–	0	–	10^{-4}	10^{-4}	0.84	0.84
	$X \sim N(\mu_k, \sigma_k^2/c_W^2)$							
0.22	0.87	–	0	–	0.03	0.03	0.09	0.10
0.35	0.78	–	0	–	0.01	0.02	0.21	0.22
0.55	0.62	–	0	–	0.005	0.006	0.37	0.38
0.73	0.44	–	0	–	0.002	0.002	0.55	0.56
0.92	0.16	–	0	–	10^{-4}	10^{-4}	0.84	0.84
	$X \sim N(\mu_k, \sigma^2/c_W^2)$							
0.22	0.94	–	0	–	10^{-4}	10^{-4}	0.06	0.07
0.35	0.77	–	0	–	10^{-5}	10^{-5}	0.23	0.24
0.55	0.45	–	0	–	10^{-6}	10^{-6}	0.55	0.56
0.73	0.17	–	0	–	10^{-7}	10^{-7}	0.83	0.84
0.92	0.01	–	0	–	10^{-9}	10^{-9}	0.99	0.99

3.4 *Simulation studies of information sources*

I extend simulations in Sheppard and Prentice (1995) to examine the association between dietary fat and cancer. Briefly, assume the model (3.2) with a single exposure and $\alpha = -6.079$, $\beta = \beta_W = \beta_B = 0.002937$. The distribution of x is fixed across $K = 21$ groups with group-specific means $\overline{x}_k$ and $\sigma_B^2 = \text{var}(\overline{x}_k)$. The within-group covariate distribution varies both in terms of the scaling of the variance ($\sigma_{Wk}^2 = (\sigma_k/c_W)^2$, $c_W = \{1/2, 1, 2, 4, 16\}$) and whether σ_{Wk}^2 varies with the group mean or is constant across groups (lognormal, normal varying mean, normal constant mean). The overall exposure variance is summarized as $\sigma_T^2 = \sigma_B^2 + \overline{\sigma_W^2/c_W}$. The random group effect (e^{v_k}) is a gamma-distributed random variable with mean 1 and variance 0.0476. Ten thousand exposures and binary outcomes are simulated with events determined by the individual-level model mean (2.1) for an average of 19 (range 5–50) events per group. There are 500 replicates.

Table 2 gives the sources of information (relative to the total information from an individual-level analysis) for both the individual-level and aggregate analyses. The total information is the average total information summed over both parameters in the individual-level model (i.e. all diagonal cells in $I_I(\beta)$ of (3.4)). The dashes in the table indicate quantities that don't exist (see (3.4)). Assuming $\beta_W = \beta_B$ demonstrates the biologic inference setting.

The relative amount of information in the between versus within area components shifts as a function of the relative variation in the exposure distribution (see σ_B^2/σ_T^2). As expected, in the individual-level study, the within-area component of information contributes most to the total information when the exposure varies more within areas. The individual and aggregate information estimates are nearly identical in the between-group component of information for both β_W and β_B. For all distributions, the information

for β_W in the between groups component is small, regardless of the analysis approach. It is negligible for the normally distributed predictor with constant variance. It is larger when the mean and variance of x are correlated within areas and largest when that distribution is also skewed. In the most extreme case of a lognormal exposure with most of the exposure variation within groups (first row of the table), the information in the between-group component is equal for β_W and β_B. This is the situation where reliance on a group-level analysis for estimation of β is least desirable. There is very little information in the between-group component and half the total information in the group-level analysis comes from high-order differences of group means. This piece will be extremely difficult to estimate in practice and impossible when only summary statistics such as area means are available, as is common in ecological studies.

3.5 *Model specification bias in ecological studies relative to aggregate data studies*

The ecological model (2.4) is mis-specified whenever the disease model is not linear and $\bar{\mathbf{x}}_k \neq \mathbf{x}_{ki} \forall i, k$. The specification bias for $\hat{\beta}$ in the ecological model can be quantified with estimating equations. The score equation for the ecological model will be biased because a mis-specified model is being fit. I correct the ecological estimating equation to make it unbiased and from this derive the bias to $\boldsymbol{\beta}$, the health effect parameter of interest.

The mean model parameters to be estimated are $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta})$. Using the notation from Section 3.1, write the score for the ecological model as

$$U(E, \boldsymbol{\theta}) = \sum_{k=1}^{K} D_k^T(E) V_k^{-1}(E)(\bar{y}_k - \mu_k(E))$$

where $D_k(E) = \partial \mu_k(E)/\partial \boldsymbol{\theta}$. Because $\mathrm{E}(\bar{y}_k) \neq \mu_k(E)$, $\mathrm{E}(U(E, \boldsymbol{\theta})) = -b(\boldsymbol{\theta})$. The unbiased estimating equation is

$$U^*(\boldsymbol{\theta}) = \sum_{k=1}^{K} D_k^T(E) V_k^{-1}(E)(\bar{y}_k - \mu_k(E) + \mu_k(E) - \mu_k(A)) = U(E, \boldsymbol{\theta}) + b(\boldsymbol{\theta}),$$

where

$$b(\boldsymbol{\theta}) = \sum_{k=1}^{K} D_k^T(E) V_k^{-1}(E) \mu_k(E) \left[1 - \frac{1}{n_k} \sum_{i=1}^{n_k} \exp\{(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)\boldsymbol{\beta}\} \right]. \tag{3.5}$$

Assume $\hat{\boldsymbol{\theta}}_b$ is the solution to $U(E, \boldsymbol{\theta}) = 0$. From a first-order Taylor series expansion of $U(E, \hat{\boldsymbol{\theta}}_b)$ about $\hat{\boldsymbol{\theta}}_b = \boldsymbol{\theta}$, the estimate of $\boldsymbol{\theta}$ from using the mis-specified model will be biased from $\boldsymbol{\theta}$ by $-I(E, \boldsymbol{\theta})^{-1} b(\boldsymbol{\theta})$ where $I(E, \boldsymbol{\theta}) = -\mathrm{E}(U'(E, \boldsymbol{\theta}))$. The bias is present even though the model has the correct link function. (Additional bias may result from the common practice of using a different functional form for the ecological model.) The term within square brackets in (3.5) determines the degree of bias. Clearly when $\mathbf{x}_{ki} = \bar{\mathbf{x}}_k \forall k, i$ there is no bias. Furthermore, when $n_k^{-1} \sum_{i=1}^{n_k} \exp\{(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)\boldsymbol{\beta}\}$ is constant across groups, the bias is absorbed into the intercept parameter α and doesn't affect $\boldsymbol{\beta}$. If this term is correlated with $\bar{\mathbf{x}}_k$, the estimate of $\boldsymbol{\beta}$ will be biased. (For instance, assume $n_k^{-1} \sum_{i=1}^{n_k} \exp\{(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)\boldsymbol{\beta}\} = \exp(\rho \bar{\mathbf{x}}_k \boldsymbol{\beta})$ and note that $U^*(\boldsymbol{\theta})$ has parameters $\boldsymbol{\theta} = (\alpha, (\rho + 1)\boldsymbol{\beta})$).

Correction for some or all of this bias can be accomplished by incorporating additional predictors in the ecological regression that summarize the joint distribution of $\mathbf{x}$. For instance, when $\mathbf{x}$ has a normal distribution, inclusion of the (known) variance of $\mathbf{x}$ alone will remove all bias:

$$\mu_k^*(E) = \mu_k(E) \exp(\boldsymbol{\beta}^T \mathrm{var}_k(\mathbf{x})\boldsymbol{\beta}/2) \tag{3.6}$$

Table 3. *Exposure effect estimates ($\times 10^{-3}$) under various group-level models and exposure distributions*

Exposure			Ecological Model % Bias		Ecological		Var. Adj. Ecological		Aggregate	
W/in-Group Distribution	W/in-Group Variance	$\frac{\sigma_B^2}{\sigma_T^2}$	Est.	Actual	$\hat{\beta}$	SE($\hat{\beta}$)	$\hat{\beta}$	SE($\hat{\beta}$)	$\hat{\beta}$	SE($\hat{\beta}$)
Lognormal		0.35	19	17	3.433	0.540	3.028	0.424	2.933	0.384
		0.73	3	3	3.040	0.594	2.951	0.561	2.947	0.558
Normal	varies	0.35	14	13	3.312	0.547	2.930	0.434	2.930	0.434
		0.73	3	3	3.030	0.590	2.943	0.557	2.943	0.557
Normal	constant	0.35	0	0	2.953	0.582	2.952	0.582	2.953	0.582
		0.73	0	0	2.943	0.606	2.943	0.606	2.943	0.606

(Richardson *et al.*, 1987; Plummer and Clayton, 1996). Inspection of $\mu_k^*(E)$ indicates that the bias using $\mu_k(E)$ will be larger the more $\text{var}_k(\mathbf{x})$ increases with $\mathbf{x}$, for larger $\boldsymbol{\beta}$, or with increased convexity of the link function ($\exp(\cdot)$ here) (Wakefield and Salway, 2001).

3.6 *Specification bias studies*

I illustrate the impact of specification bias by simulation using the same example. To facilitate comparison with ecological studies as they would tend to be analyzed, I do not include any group-specific frailties. Otherwise, binary outcomes are simulated based on (2.1).

Table 3 gives results for six scenarios using the three covariate distributions and two ratios of exposure variation, σ_B^2/σ_T^2. (These ratios bracket estimates previously reported for international diet and cancer studies (Prentice and Sheppard, 1995).) I report only β since this is the parameter of interest for biologic inference. Estimates are given for the percent bias in an ecological model, followed by the parameter estimates and standard errors in the ecological model, the variance-adjusted ecological model (replace $\mu_k(E)$ from (2.4) with $\mu_k^*(E)$ in (3.6)), and the aggregate model.

As discussed in the previous section, the ecological estimate of β is unbiased under a normally distributed covariate with constant variance across groups since the bias is absorbed into the intercept (see also (Plummer and Clayton, 1996)). Related to this point, the between-group component of information has no information on β_W (Table 2). For the other two exposure distributions, the ecological regression is positively biased because the group means are positively correlated with the within-group variances and the between-group component of information contributes information on β_W (Table 2). In all cases the estimated bias is close to the actual bias. Their difference is due to the inaccuracy in the single-step Taylor series expansion. There is more bias in the lognormal distribution because the higher-order moments are also correlated with the mean. The bias is larger when there is a larger proportion of within-group variation in the exposure (compare $\frac{\sigma_B^2}{\sigma_T^2} = 0.35$ versus 0.73). The variance adjustment is sufficient to remove the model specification bias for the normally distributed predictor because (3.6) completely characterizes the bias and I estimate the within-area variance from the entire population of 10 000 exposures. Wakefield and Salway (2001) show poorer performance of the variance-adjusted ecological model when the variances are estimated from small samples.

4. Discussion

This paper highlights three major points that should be considered for any group-level study. First, a group-level analysis does not use all the information in the data. Depending upon the relative amount of variation in the exposure between versus within groups, there can be considerable information lost by aggregating over individuals. Second, inference about the exposure effects from a group-level study relies on the between-group parameter, β_B, which could have a different scientific interpretation from the within-group parameter, β_W. Inference from group-level studies is cross-level in the sense that they give information almost exclusively about β_B, but for biologic inference the parameter of interest is β_W. Finally, since almost completely separate components of variation in the covariate are used to estimate β_W versus β_B, each could be subject to different biases. In a study of individuals with exposures that vary both within and between groups, this could be used to advantage when trying to identify and control sources of bias.

While I've focused primarily on bias and information related to exposure variables, the ability to partition information applies to all covariates and can be used to inform questions about the impact of potential confounders and mis-measured covariates in both group-level and clustered data studies. For instance, in environmental epidemiology studies it may be reasonable to assume cross-level inference will be valid for the environmental exposure of interest, while it will be more plausible to assume non-environmental covariates will have different within- and between-cluster effects on the outcome. In particular, behaviorally derived covariates will often have different meaning and thus different health effects between versus within clusters.

In any specific application it is crucial to gain in-depth understanding of the exposure of interest and its distributional properties. As an example, Sheppard and Prentice (1995) examine the impact of exposure measurement error and uncontrolled confounding in dietary studies analyzed at the group versus individual level. Disconnection between available and ideal exposure data can lead to substantial bias in exposure effect parameters, regardless of the design. In semi-individual and ecological studies, specification bias is likely. Often incorporation of information about within-group covariate distributions will be necessary, although in some applications bias from omitting this information may be small. Individual studies should always suspect classical measurement error. Residual confounding cannot ever be ruled out in an observational study and thus must be considered in all designs.

While the issues of cross-level bias and variation are critical to the validity of group-level studies, they are not unique to them. They are present in *all* multilevel studies. Although not the focus of this work, I note that cross-level bias can be hidden in covariate effect estimates from longitudinal studies whenever proper precautions are not taken in the analysis. Furthermore, even when the between- and within-cluster effects are equal, the problems inherent in correctly incorporating multilevel information in an analysis are present whenever covariates vary both within and between clusters. Biases due to measurement error or confounding can dominate just one of the effect estimates, either the between- or the within-group effect, because of its impact on the exposure at just one of the two levels of information. For instance, measurement error may bias the within-group effect estimate while leaving the between-group effect estimate unchanged. Likewise, a confounding variable may be associated with an exposure variable at only one level, either within or between groups, thus leaving the exposure effect estimate at the other level of analysis unbiased. As an example, typical air pollution panel studies collect repeat pollutant and outcome measures over time in a sample of individuals. Analyses that separate the within- and between-subject effects can obtain very different estimates for these effects. Yu *et al.* (2000) were faced with this difficulty in an analysis of the role of pollutants, specifically carbon monoxide and particulate matter, on symptoms of asthma in the greater Seattle area. Children in this study were observed for about two months each over a two-year period. The exposure measurements were derived from ambient monitors, so all the differences in subject-specific exposure means were due to the subject-specific time periods of

observation. However, season is a strong determinant of both air pollution and asthma symptoms. Since residual seasonal confounding in the between-subject exposure effects was likely, this paper only reported the within-subject exposure effects.

ACKNOWLEDGEMENTS

This work was supported by grant ES08062-03 from the National Institute of Environmental Health Sciences, NIH. The contents are solely the responsibility of the author and do not necessarily represent the official views of NIEHS. I wish to thank Jon Wakefield and Katherine Guthrie for useful discussions, and Katherine Guthrie for assistance with the simulations.

REFERENCES

ARMSTRONG, B. AND DOLL, R. (1975). Environmental factors and cancer incidence and mortality in different countries, with special reference to dietary practices. *International Journal of Cancer* **15**, 617–631.

BESAG, J., YORK, J. AND MOLLIE, A. (1991). Bayesian image restoration with two applications in spatial statistics. *Ann. Inst. Statist. Math.* **43**, 1–59.

DOCKERY, D. W., POPE, C. A., XU, X., SPENGLER, J. D., WARE, J. H., FAY, M. E., FERRIS, B. G. AND SPEIZER, F. E. (1993). An association between air pollution and mortality in six U.S. cities. *New England Journal of Medicine* **329**, 1753–1759.

DOLL, R. AND PETO, R. (1981). The causes of cancer: quantitative estimates of avoidable risks of cancer in the United States today. *Journal of the National Cancer Institute* **66**, 1196–1265.

EPA (1996). *Air Quality Criteria for Particulate Matter, Volume III*. Research Triangle Park, NC: Office of Research and Development, National Center for Environmental Assessment, US Environmental Protection Agency.

FIREBAUGH, G. (1978). A Rule for Inferring Individual-Level Realtionships from Aggregate Data. *American Sociaological Review* **43**, 557–572.

GREENLAND, S. (1992). Divergent biases in ecologic and individual-level studies. *Journal of the National Cancer Institute* **11**, 1209–1223.

GREENLAND, S. (2001). Ecologic versus individual-level sources of bias in ecologic estimates of contextual health effects. *International Journal of Epidemiology* **30**, 1343–1350.

GREENLAND, S. (2002). A review of multilevel model theory for ecologic analyses. *Statistics in Medicine* **21**, 389–395.

GREENLAND, S. AND MORGENSTERN, H. (1989). Ecological bias, confounding, and effect modification. *International Journal of Epidemiology* **18**, 269–274.

GREENLAND, S. AND ROBINS, J. (1994). Ecological studies—Biases and misconceptions and counter-examples. *American Journal of Epidemiology* **139**, 747–760.

GUTHRIE, K. A. AND SHEPPARD, L. (2001). Overcoming biases and misconceptions in ecological studies. *Journal of the Royal Statistical Society, Series A* **164**, 141–154.

HAMMOND, J. L. (1973). Two sources of error in ecological correlations. *American Sociological Review* **38**, 764–777.

HOMLES, M. D., HUNTER, D. J., COLDITZ, G. A., STAMPFER, M. J., HANKINSON, S. E., SPEIZER, F. E., ROSNER, B. AND WILLETT, W. C. (1999). Association of dietary intake of fat and fatty acids with risk of breast cancer. *Journal of the American Medical Association* **281**, 914–920.

KUNZLI, N. AND TAGER, I. B. (1997). The semi-individual study in air pollution epidemiology: a valid design as compared to ecologic studies. *Environmental Health Perspectives* **105**, 1078–1083.

LANGBEIN, L. I. AND LICHTMAN, A. J. (1978). *Ecological Inference*. Beverly Hills, CA: Sage Publications.

MORGENSTERN, H. (1982). Uses of ecologic analysis in epidemiologic research. *American Journal of Public Health* **72**, 1336–1344.

MORGENSTERN, H. (1998). Encyclopedia of Biostatistics. Armitage, P. and Colton, T. (eds), *Ecologic Study*. New York: Wiley, pp. 1255–1276.

MORGENSTERN, H. AND THOMAS, D. (1993). Principles of Study Design in Environmental Epidemiology. *Environmental Health Perspectives Supplement* **101**, 23–38.

NEUHAUS, J. M. AND KALBFLEISCH, J. D. (1998). Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics* **54**, 638–645.

PALTA, M., YAO, T.-J. AND VELU, R. (1994). Testing for omitted variables and non-linearity in regression models for longitudinal data. *Statistics in Medicine* **13**, 2219–2231.

PIANTADOSI, S., BYAR, D. P. AND GREEN, S. B. (1988). The ecological fallacy. *American Journal of Epidemiology* **127**, 893–904.

PLUMMER, M. AND CLAYTON, D. (1996). Estimation of population exposure in ecological studies. *Journal of the Royal Statistical Society, Series B* **58**, 113–126.

POPE, C. A., BURNETT, R. T., THUN, M. J., CALLE, E. E., DREWSKI, D., ITO, K. AND THURSTON, G. D. (2002). Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *Journal of the American Medical Association* **287**, 1132–1141.

PRENTICE, R. L. (1996). Measurement error and results from analytic epidemiology: dietary fat and breast cancer. *Journal of the National Cancer Institute* **88**, 1738–1747.

PRENTICE, R. L., KAKAR, F., HURSTING, S., SHEPPARD, L., KLEIN, R. AND KUSHI, L. H. (1988). Aspects of the rational for the Women's Health Trial. *Journal of the National Cancer Institute* **80**, 802–814.

PRENTICE, R. L. AND SHEPPARD, L. (1990). Dietary fat and cancer: consistency of the epdemiologic data, and disease prevention that may follow from a practical reduction in fat consumption. *Cancer Causes and Control* **1**, 81–97.

PRENTICE, R. L. AND SHEPPARD, L. (1995). Aggregate data studies of disease risk factors. *Biometrika* **82**, 113–125.

RICHARDSON, S., STUCKER, I. AND HEMON, H. (1987). Comparison of relative risks obtained in ecological and individual studies: some methodological considerations. *International Journal of Epidemiology* **16**, 111–120.

SCHWARTZ, J. E. AND STONE, A. A. (1998). Strategies for analyzing rcological momentary sssessment data. *Health Psychology* **17**, 6–16.

SCOTT, A. J. AND HOLT, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association* **77**, 848–854.

SHEPPARD, L., LEVY, D., NORRIS, G., LARSON, T. V. AND KOENIG, J. Q. (1999). Effects of ambient air pollution on nonelderly asthma hospital admissions in Seattle, Washington, 1987–1994. *Epidemiology* **10**, 23–30.

SHEPPARD, L. AND PRENTICE, R. L. (1995). On the reliability and precision of within- and between population estimates of relative rate parameters. *Biometrics* **51**, 853–863.

WAKEFIELD, J. AND SALWAY, R. (2001). A statistical framework for ecological and aggregate studies. *Journal of the Royal Statistical Society, Series A* **164**, 119–137.

WILLETT, W. C. (2001). Diet and cancer: one view at the start of the millennium. *Cancer Epidemiology, Biomarkers and Prevention* **10**, 3–8.

WILLETT, W. C. AND STAMPFER, M. J. (1990). Dietary fat and cancer: a comeback for ecological studies? *Cancer Causes and Control* **1**, 101–102.

YU, O., SHEPPARD, L., LUMLEY, T., KOENIG, J. Q. AND SHAPIRO, G. G. (2000). Effects of ambient air pollution on symptoms of asthma in Seattle-area children enrolled in the CAMP study. *Environmental Health Perspectives* **108**, 1209–1214.

ZOCK, P. L. (2001). Dietary fats and cancer. *Current Opinion in Lipidology* **12**, 5–10.

[*Received February 7, 2001; first revision May 24, 2002; second revision July 3, 2002; accepted for publication July 22, 2002*]